THE ONTARIO INSTITUTE FOR STUDIES

OISE CIR
3 0005 03037 6217
150.8 A673 no.154 c.1
Woodworth, Robert Sessions.
Archives of psychology --
R.W.B. JACKSON LIBRARY

Measuring Teaching Efficiency Among College Instructors

BY

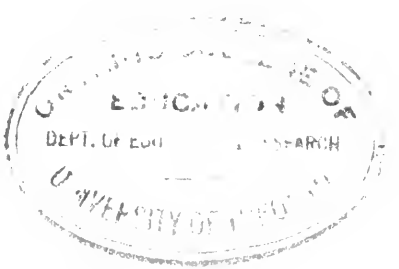
GEORGE W. HARTMANN, PH.D.

Department of Psychology, Pennsylvania State College

ARCHIVES OF PSYCHOLOGY

R. S. WOODWORTH, EDITOR

No. 154



150.8
A673
no. 154
1933

NEW YORK

July, 1933

ARCHIVES OF PSYCHOLOGY
COLUMBIA UNIVERSITY, NEW YORK CITY

The Subscription price is six dollars per volume of about 500 pages. Volume I comprises Nos. 2–10; Volume II, Nos. 11–18; Volume III, Nos. 19–25; Volume IV, Nos. 26–32; Volume V, Nos. 33–39; Volume VI, Nos. 40–46; Volume VII, Nos. 47–52;

THE LIBRARY

The Ontario Institute

for Studies in Education

Toronto, Canada

38. Effect of Humidity on Nervousness and on General Efficiency: L. I. STECHER. 90c. (Cl., $1.15.)

78. Aggressive Behavior in a Small Social Group: E. M. RIDDLE. $1.75.

(Continued on inside back cover.)

Measuring Teaching Efficiency Among College Instructors

JUN 3 1969
FOR STUDIES IN EDUCATION

BY

GEORGE W. HARTMANN, PH.D.
Department of Psychology, Pennsylvania State College

ARCHIVES OF PSYCHOLOGY
R. S. WOODWORTH, EDITOR
No. 154

UNIVERSITY OF TORONTO

NEW YORK
July, 1933

TABLE OF CONTENTS

CHAPTER		PAGE
I.	Introduction and conceptual orientation	5
II.	Analysis of previous attempts to determine teacher efficiency	10
III.	The experimental procedure: description of tests	20
IV.	Results and discussion	29
V.	The effect of formal school instruction upon personality organization	39
	A. Changes in various attitudes resulting from college courses	39
	B. Changes in the mean valuational types of students in different subjects and under different instructors	40
VI.	Conclusions	44
VII.	Summary	45

MEASURING TEACHING EFFICIENCY AMONG COLLEGE INSTRUCTORS

CHAPTER I

INTRODUCTION AND CONCEPTUAL ORIENTATION

The "effectiveness of instruction" is a professional phrase commonly employed to designate two distinct things. It may refer to the superiority of a certain teaching method or procedure over another or it may indicate the amount of teaching skill possessed by one individual as contrasted with others. Both uses postulate the existence of at least one index, mark, or criterion, by means of which the relative excellence of a device or a person may be ascertained. The reasons for selecting some signs rather than others as symptomatic of various degrees of competence are generally obscure, and the justification of the choice normally rests upon social or philosophical grounds which the positive scientist accepts as ultimate, even though he may be fully aware of their tentative character.

The technician, *qua* technician, who proposes to develop or improve a way of measuring teaching efficiency, does not ask if it be worth while to do so. Contemporary man seems to be irrevocably committed to the ideal of perpetual progress; he is constantly seeking improved ways of bettering himself, his activities, his institutions, and the natural world in which he finds himself. Consequently, those who are entrusted with the function of transmitting the social heritage of the race are interested in any procedure which will aid in the discharge of that responsibility, whether it consists in a new technique of presenting subject-matter or in a more objective way of appraising merit in the teacher himself. Since it is the latter problem which will be our primary concern in this study, we must remind ourselves why this is an educational desideratum. Just as we use marks or scores to measure the quality of a student's performance, so we need estimates of a teacher's skill to aid in the process of recruiting, rewarding, promoting and maintaining a satisfactory instructional staff. Wholly apart from the matter of incentive and administrative control, every progressive teacher would be curious and happy to know of some convincingly objective mode of revealing his true rank in the company of his fellows.

The efficiency of any member of a profession is measured in terms suitable to that career. Statistically, one may suppose that

a physician's competence would be settled by the proportion of cures to the number of cases treated, the comparative difficulty of the therapy exercised, etc.; and a similar situation would obtain for the practice of law. To some extent, a politician's "efficiency" may be judged by the votes he receives or by the judgment of historians and posterity. Likewise, a military man's efficiency is roughly measurable by the number and decisiveness of his victories; the artist's by the enthusiastic appreciation of people of taste; the factory workman's by the piece-rate wages he receives, etc. But all these methods are distressingly inaccurate and so distorted by the operation of favoritism, chance, and special circumstances that the scientifically-minded can consider them only because nothing better is available. It is to the glory of the teaching profession that many of its leaders are unremittingly seeking better instruments for classifying and arranging the variations in skill which are known to exist but which have proved so resistant to quantitative attack.[1]

A grave difficulty with past endeavors to accomplish this result has been the lack of confidence of the teachers in the "fairness" of the widely-used rating instruments, with the consequent destruction of morale when it is made the basis of important modifications in professional status. Much of the unfavorable criticism which has been levelled at the rating scale has come from prejudiced and incompetent observers, but experts recognize the justice of protests when the necessary conditions of adequate knowledge, training, and

---

[1] Since an economic note invariably runs through all contemporary discussions of efficiency, there is a certain appropriateness in securing an economist's opinion of the matter. "It is elementary that wages can be an effective incentive only when the compensation of each employee depends upon his performance. This, of course, presupposes that the management can measure what each worker does. The performance of workers includes more than is usually supposed. The quantity and quality of output are the two most important elements. In addition to these, however, performance includes many other things —such as how economical each worker is of materials, power, and light, the amount of spoiled work for which he is responsible, his breakage of equipment, the rate at which he wears out equipment, his ability and willingness to suggest improvements in methods and products, his regularity of attendance, and the number of jobs which he is capable of doing.

"The measurement of these elements of efficiency is often extremely difficult or positively impracticable. The volume of output is usually the easiest but even it cannot always be ascertained. In the case of non-repetitive work, for example, such as special order and repair work, output often cannot be measured because there is no common unit of performance. In other cases, measurement is impossible because the contributions of individual workmen become indistinguishably merged in a common result. This is illustrated by many assembling operations, by construction work, maintenance of railway track, and the work of yard gangs."—S. H. Slichter, *Modern Economic Society*, New York, Holt, 1931, pp. 594–595. It would be gratuitous to point out the many analogies to teaching situations which the preceding excerpt reveals.

independence of judgment on the part of several judges are not met. Advocates of a more democratic type of control in the educational structure have favored an extension of the rating scheme to embrace judgments of colleagues and pupils rather than restricting the privilege of estimating talent to those in authority. While this is undoubtedly a commendable experimental trend, the emotional disturbances which frequently ensue upon the wounding of one's *amour propre* have increased the ardor of the search for a more satisfying and intellectually acceptable measuring rod.

If we interpret our problem as that of gauging the efficiency of a teaching personality by the relative position he occupies on some common scale of merit, one of our first duties is to clarify the concept of efficiency in general. Fortunately, we may profit here from the experience of personnel chiefs and other experts in scientific management in business and industry. Historically, the first terminological struggles were made by engineers when they selected the ratio between the energy intake and the output of a machine as a measure of its efficiency. Can the same viewpoint be adopted unaltered in the human realm? Applied psychologists very soon discovered that significant modifications in the engineering notion were needed before it could be made serviceable in dealing with human performance. Since educational activities are so intrinsically personal, it is impossible to adopt the simple view that the most efficient process is the one that brings the maximum return for the least expenditure of energy. While the complications introduced by the "human element" are great, they are not sufficiently disturbing to break down the mathematical or biological fitness of the customary equation. As modified by Poffenberger,[2] the following statement appears to give the analytical guidance required: "The ideal of human efficiency would be the production of the *maximum output* of the *highest quality* in the *shortest time,* with the *least expenditure of energy* and with the *maximum satisfaction.*"

How can this be applied to the teaching situation? By substituting those corresponding terms which will preserve the pattern of the formula and yet do justice to the concrete variations in the activity. The "output" of the teacher is clearly the character of the changes produced by him in the learner, as Thorndike observed a quarter-century ago. The oral and written behavior of the teacher, however strenuous, is meaningful only with reference to this goal. Of course, these changes must be "socially desirable"

[2] A. T. Poffenberger, *Applied Psychology,* New York, Appleton, 1927, pp. 349–352.

before effective teaching can be said to have occurred—a feature which corresponds to the item of "quality" in Poffenberger's account. The teacher who produces the most valuable alterations in the pupils' natures is, other things being equal, the most efficient. But, as both wits and fools have repeatedly warned, other things are seldom equal. Time is of the essence of living and that instructor who accomplishes certain results in a semester is more efficient than the one who must use the whole year. Unfortunately, this still does not do full justice to the subjective aspect of the equation, for the teacher who must sweat blood himself and drive his charges unmercifully to attain a given standard may be guilty of unwarranted demands upon the metabolic resources of the organism. This item leads naturally to the last consideration involving pleasure or happiness, an intangible one, to be sure, but none the less real: The efficient teacher is also one, who in addition to making equal modifications in equal time shows a surplus of contentment or satisfaction for both himself and his students. If we combine and condense these separate aspects, our final definition will then read: "*The ideally efficient teacher is the one who can accomplish the largest number of important and socially desirable changes in the greatest number of pupils in the shortest possible time, with the least expenditure of energy and with the maximum satisfaction in the learning process and its outcomes by all concerned.*"

Externally, the purposes of scientific research are beautifully served by this comprehensive formula. It lends itself readily to the experimental technique of the single variable and multiple constants. The component factors can be isolated and controlled about as readily as they can be regulated in any complex social situation, and as every reader who has followed the development of the discussion to this point can see, the statement is easily converted into a form applicable to the measurement of the efficiency of any *method* as well as of any *individual*. The chief limitation of this standpoint is its disregard of the probability that a change in any part not only influences other parts but has marked repercussive effects upon the total in which they are comprised. As the Gestalt theorists have recently cautioned us, we must avoid atomistic thinking in fields where it is unwarranted.

The major analytical error which one is likely to commit is the assumption that the merits of any teaching method are independent of the caliber of the person using it. So far is this from being true that there is strong reason for believing that a "poor" method (*i.e.*, one inferior on psychological and statistical grounds) in the hands

of a "good" instructor is a better teaching risk than a demonstrated "good" method employed by a "poor" instructor. Failure to appreciate the significance of this nice organic adaptation of the workman to his tool has resulted in much futile criticism of the lecture procedure on the college level. When used by a person to whom it is ill-suited—*e.g.,* some shy scholar with a feeble voice and an unimpressive physique—then of course it is a torment which no one can sanction. But when wielded by a master of expository skill with an appropriate "platform manner," it is a powerful pedagogical instrument, as the distinguished record of many lecturers in the German universities testifies. Educationists who have been blind to this fact are reproducing the persistent and pardonable error of the early vocational counsellors when they thought in terms of the evils of "square pegs in round holes," a view which we now know to be faulty because it posited a static relationship between the man and the job which never obtains. Instead, a reciprocal modification occurs. One may say that square pegs become a bit roundish and the round holes assume squarish outlines; at any rate, the worker-in-his-work[3] becomes the true indissoluble functional unit with which vocational guidance must operate. Similarly, the teacher-and-his-method may be the new dynamic organization to be introduced into our studies of classroom technique.

There are other considerations which make it plain that excessive concern with pure method as such is an artificial and unprofitable occupation. The "master method" has not yet been found because the relative merit of any procedure appears to vary not only with the teacher as indicated above, but with the subject of instruction, and with the mental level and other pertinent traits of the pupils as well. In fact, I suspect that the teacher-method unit just mentioned is still too atomistic to reflect the actual degree of integration which may be better represented by the teacher-method-subject-pupil configuration. It is highly probable that the typical parallel-group experiment is engaged in comparing such unified wholes or gestalten rather than the single variables which we have constructed mentally.

Equipped with this stock of ideas and the attitude of relentless self-criticism which it enjoins, we shall be the better prepared to avoid the subtle dangers of philosophical innocence all too evident in many of the "exact" investigations of contemporary science.

---

[3] For a persuasive exposition of this point of view, see W. D. Scott, R. C. Clothier, and S. B. Mathewson, *Personnel Management,* revised edition, McGraw-Hill, New York, 1931, pp. 8–18.

CHAPTER II

ANALYSIS OF PREVIOUS ATTEMPTS TO DETERMINE TEACHER EFFICIENCY

One of the saddest commentaries upon the character of human knowledge is that the most precise, detailed, and reliable information pertains to areas which seem to possess relatively little intrinsic worth to the mass of mankind, while those fields which touch us most intimately are just the ones in which the densest ignorance prevails. Compare, for instance, the number, variety, and certainty of the facts available on the nature of color discrimination among invertebrates with the quality of our information on the higher thought processes in man; or the difference between our understanding of the Albigensian heresy and the causes and cure of the present economic depression!

The same pathetic contrast besets us in the educational sphere. We know with great exactness how Johnny's eyes wiggle when leaping over one of the Rover Boy stories and how they are fixated in following a page of algebraic equations, but what happens in his nervous system a few years later when he comprehends the futility of high tariffs is still a matter for lively debate. Our "cobblestone-counting" type of educational research proclaims with extraordinary assurance what proportion of high-school chemistry text-books is devoted to the household arts while the quest for a more persuasive way of establishing a dependable set of ideals in our youth is blocked by the obstacles which nature and our own obtuseness have placed in the path.

This viewpoint may be erroneously founded upon some odd illusory effect which the progress of science creates, but as an opinion it is only strengthened by the recency and paucity of sound endeavors to deal with the subject of this investigation. Apparently until very recent decades, the best minds were engaged with other problems, despite the general recognition that the larger interests of education demanded other means of identifying variations in teaching efficiency than the defective ones currently employed. Contemporary students of the matter are, therefore, all the more indebted to those pioneers who made the first efforts in this direction.

Most of these early studies appear to have been motivated less by an interest in the measurement of teaching efficiency as such than by a desire to predict teaching success from a knowledge of

other relevant facts about the individual. Clearly one had to determine who the "successful" or "unsuccessful" teachers were before the prognostic devices could be applied. Good and poor teachers must be identified—but how? Competence on the job seemed the sole valid criterion and measures of it could be obtained only in the form of ratings from supervisors, principals, and other administrative officers. We now know that this was the weakest element in the whole research structure. Most of these promising and ambitious investigations are characterized by stupendous industry in collecting mass data and high skill in applying the multiple correlation technique to them—but the findings are generally disappointing. Why? Not so much because the prognostic instruments are faulty; nor because statistical certainty is lowered by operating with a select group rather than with the general population; nor because many of the qualities correlated are truly irrelevant (although all these explanations contribute something);—but largely, I believe, because the authorities could not supply a dependable classification of their personnel. Under favorable conditions, the best rating scales constructed by applied psychologists are reasonably consistent and reliable, but in the absence of independent standards of success, their validity remains unknown.

Some such interpretation seemingly must be allowed for if the fantastic finding that teaching ability is not associated with scholarship as measured by academic grades,[1] professional marks,[2] or intelligence is to be understood.[3] One's faith in the reign of law in the human and social order would be shattered were these conclusions to be accepted at their face value, and it is only partly restored by the discovery of the modest relationship between cadet teaching grades and later teaching success.[4] Even specifically designed aptitude tests reveal no more satisfactory connection. The most plausible explanation is that the ranking of teachers by their overlords must be subject to serious errors. If twenty administra-

---

[1] F. B. Knight, *Qualities related to success in teaching*, New York, Teachers College, Columbia University, Contributions to Education, No. 120, 1922, p. xiii. His r was 0.153. It is interesting ot note that the first number in this distinguished series is directly concerned with this topic. Cf. J. L. Meriam, *Normal school education and efficiency in teaching* (1906).

[2] F. L. Whitney, *The prediction of teaching success*, J. Ed. Res. Mon., No. 6, 1924, Public School Publishing Co., Bloomington, Ill. His obtained r was 0.143 (see p. 20).

[3] W. H. Pyle, *The relation between intelligence and teaching success*, Ed. Adm. & Sup., 1928, 14, 257–267. For 99 cases, the r after one year of public school teaching was 0.034; after two years, 0.023.

[4] Roy R. Hillman, *The prognostic value of certain factors related to teaching success.* Ashland, O., A. L. Garber Co., 1931. His r was 0.36.

tive judges agree in placing teacher A at the top of a group of one hundred and teacher B at the bottom, it is quite likely that this pooled verdict indicates the "true" position of these individuals. But every one realizes that this is not the usual situation. The "halo" effect is a constant source of trouble, and instead of emanating from twenty judges, most ratings come from one—where more are obtained, disagreement rather than unanimity in placement is the rule. Hence, while one may appreciate the defensibility of resorting to the opinion of competent judges where objective quantitative measures cannot be applied, the limitations of this procedure must always be recognized.

At this stage, it may be wise to admit that the possibility of having followed a wrong clue is very real. Consequently, an alternative approach through the utilization of achievement tests has been suggested, which commends itself especially to those who desire more extensive evidence than that provided by the ordinary rating blank. The method of the achievement test is a more awkward and laborious tool to apply but it possesses the distinct advantage of being free from the evils of "impressionism." Its use is simple: Two instructors teaching the same subject, the same type of pupil, and operating under essentially equivalent conditions, administer the same pre-test and the same final test to their classes. Members in each group are paired or matched for initial ability on the basis of the first test records; therefore, the group which shows the higher average final score must have made the greater gain in subject-matter mastery.[5] It is a reasonable assumption to attribute this difference to the greater skill or efficiency of the teacher involved. How he accomplished this result is, of course, a matter for supplementary examination. With proper precautions exercised by an experimenter trained in the technique of scientific research, valuable and pertinent data relating to individual differences among members of the teaching staff may thus be secured. Since it was this method which was adopted in the study soon to be described, the occasional objections which have been raised against it ought to be considered at this point.

There are those who claim that it is a perversion of a test's purpose to measure teacher achievement when its real function is to

[5] A brief and simple illustration of this procedure as applied to a comparison of the relative efficiency of two *methods* rather than individuals may be found in a minor article by the author: Economy of time in college instruction, *J. Ed. Res.*, May, 1931, pp. 404–409. Whenever it is employed without due regard for the necessities of equating pupils for ability, it becomes a vicious abuse of the principle.

determine pupil accomplishment. To be sure, the original intention of a testing program is definitely modified if used for such an end, but new uses are constantly being discovered for old instruments. The rating scale seems to have been employed in stock-judging contests long before it was applied to humans or their products! Moreover, if the principle to which we have committed ourselves is sound, viz., that the work of a teacher can be estimated only by the nature of the changes wrought in the pupils, then the test results which measure at least a sample of these alterations are a perfectly legitimate source of information. It is plain that *all* measures of a teacher's success must be indirect, *i.e.*, they apply not to the teacher as such, but to certain features of his pupils which are referred back to him as the agent evoking them. The fact that a measure is indirect, however, does not invalidate it, since practically all measurement is of that type. Variations in heat are measured indirectly by shifts in the height of a column of mercury; a mother's "efficiency" is gathered from the mental and physical fitness of her children.

A more serious shortcoming of the ordinary achievement test as a symptom of instructional competence is the narrowness of the function measured. Most observers when asked to comment upon the adequacy of our standardized new-type educational examinations lament the fact that they are restricted to sheer memory items.[6] On this basis, the best teacher would merely be one who could cram the maximum amount of raw knowledge down the students' throats. But what becomes of such important goals which must also be defined in terms of pupil behavior or reaction as: inspiration and enthusiasm for scholarship, ability to formulate reasonable generalizations, ability to plan experiments for testing hypotheses, etc.?[7] These critics rightly protest that there are other objectives of instruction in addition to the recall of information. While damaging to narrowly-conceived programs of achievement measurement, this criticism does no fundamental injury to the idea because the objection is met, in principle at least, by the addition

[6] R. W. Tyler, Measuring the results of college information, *Educational Research Bulletin*, 1932, XI, 253–260.

[7] F. D. Curtis, *Second digest of investigations in the teaching of science*, P. Blakiston's Son & Co., Philadelphia, 1931. The following statement is representative: "Experimenters have frequently evaluated the relative merits of competing methods in terms of a knowledge of subject matter over a longer or shorter period, and have ignored other goals of instruction, which may be quite as important, e.g., training in scientific attitudes, in reasoning, in correct habits of reaction to given typical situations, etc." p. 6. Kilpatrick has also recently emphasized this problem of total or aggregate instruction.

of further reliable tests of attitude, character, and personality to include just those vague but significant outcomes referred to above. The good instructor, then, is one who can produce desirable increments of growth in those aspects as well.

There are several studies which show that the determination of changes in attitude is a perfectly feasible experimental problem. Kornhauser[8] found definite evidence of a shift in the direction of increased liberalism and greater liking for scientific procedure after attendance upon an economics course. McGeoch[9] examined the effect of instruction in psychology upon scores in the Pressey X-O test of emotion and observed that "an interval of 45 days showed an inconstancy of response between the first and second givings, indicating that altered affective organization is one of the unnoticed by-products of teaching." An exploratory effort of Sturges[10] is worthy of note. In a number of class experiments, initial tests on war and peace attitudes were given, where there could be no justification for expecting that the students had these adequate attitudes at the beginning of the course of study. A repetition of these tests at the end of the term showed invariably a much closer approximation to the points of view presented in the course.

Another argument against the legitimacy of judging teaching merit by the immediate test results is presented by a student of the problem as follows: "The greatest difficulty here . . . is that the product of the teacher's effort really comes to fruition not in a week, nor in a month, nor yet in a year. The fact is some ten to twenty years must pass before the fruition is attained."[11] This implies that the great teacher of art must wait until his pupil has outstripped his master before the true quality of his instruction emerges; and the great teacher of science must likewise delay self-appraisal until his students' contributions permit an opportunity to judge of the nature of the influences which have made them possible. Unfortunately, this contention proves too much. Admittedly, some effects are delayed and may remain latent for a long time; but just what these are is unknown. Even when they do

---

[8] A. W. Kornhauser, Changes in the information and attitudes of students in an economics course, *J. Ed. Res.*, 1930, 22, 288–298.

[9] J. A. McGeoch & Marion E. Bunch, Scores in the Pressey X-O tests of emotion as influenced by courses in psychology, *J. Applied Psychol.*, 1930, *14*, 150–159.

[10] H. A. Sturges, The theory of correlation applied in studies of changing attitudes, *Amer. J. Sociology*, 1927, *33*, 269–275.

[11] C. L. Jacobs, *The relation of the teacher's education to her effectiveness*, Teachers College, Columbia University, Cont. to Ed., No. 277, p. 25.

become overt, how shall we determine the precise contribution to a student's success made by some teacher years before? The difficulties of controlled experimentation become well-nigh insurmountable where the crucial check is prolonged to that extent. Furthermore, the fact that there is a high positive association between initial and final achievement in laboratory performances involving learning and between immediate and delayed recall, as well as the fact that the quick learner is *not* the quick forgetter, make it unlikely that really significant changes are not amenable to measurement *via* the vehicle of testing after a semester or year of "exposure" to a certain teacher.

A miscellaneous collection of the shortcomings of the achievement-test method of identifying good and poor teachers is listed by Barr[12] in describing his preferences as follows: "Educational tests might have been given . . . at the beginning and end of each of several successive semesters of work in order to determine the efficiency of each teacher. . . . This, however, was not done because: (1) Tests do not measure teaching success, except when applied under carefully controlled conditions. Tests measure changes in pupils, including those resulting from native capacity, from maturity, from home environment, from increased or decreased effort, from health, from outside assistance, etc. The teacher's teaching, however important it may be, is only one factor among many. (2) The application of tests necessitated certain assumptions relative to the outcome of teaching which the investigation was not prepared to make—(a) that all of the results of teaching could be measured, (b) that the objectives assumed by the makers of tests constitute the proper goals of teaching, (c) that the tests selected for use measure fully what they purport to measure, and (d) that the tests measure reliably what they purport to measure."

These difficulties are real enough but each one needs to be examined carefully before being considered insuperable. Control groups properly constructed through matching are the remedy for most of the obstacles listed. It is strange that Barr, who is so keenly aware of the limitations of testing as the previous quotation shows, nevertheless prefers to overlook the even more pronounced deficiencies of the supervisory ratings which he used as a basis for

[12] A. S. Barr, *Characteristic differences in the teaching performance of good and poor teachers of the social studies,* Public School Publishing Co., Bloomington, Ill., 1929.

selection. One reason for the existence of intelligence tests is that within one hour they segregate individuals according to mental level in a fashion which could only be crudely made even after years of acquaintance. By virtue of their specificity, they become substitutes for vast amounts of random experience and contacts. And if one must choose between the reliability of tests and the reliability of ratings, there is little doubt as to which alternative is preferred by competent authorities. One of the least pardonable offences against the canons of science is the deliberate use of an inferior experimental technique when a better one is equally available.

A sober, critical review and appraisal of the impassé facing research in this field has been made by Symonds.[13] He notes that the high correlation which Boyce[14] found between general rating for teaching efficiency and such items as "development of pupils" (.88), "growth of pupils in subject-matter" (.87), and "attention and response of class" (.86) suggested to other workers the use of pupil achievement as a measure of teaching efficiency. This implied the adoption of the Jesuit principle that faith without works is dead. "After all, the final criterion of any activity is the results produced. Traits in the teacher are valuable as measures of teaching efficiency only when they are effective in producing desirable changes of learning in pupils." Such doctrine is both true "Westernism" and good behaviorism. Consequently, the proposal that the *accomplishment ratio* be employed as a means of estimating efficiency was inevitable. As developed by Franzen, the formula read very simply:

$$\text{Teaching efficiency} = \text{Final AR} - \text{Initial AR}$$

The theoretical appeal of this equation was irresistible for not only did it appear to provide for the initial achievement of the group but, like any good quotient, it made proper allowance for differences in potential ability. Crabbs was the first to use this device in measuring the effectiveness of staff members in the elementary school, but the results were grossly disappointing since the correlations with supervisors' estimates were either miserably low, zero, or even negative. Crabbs[15] interpreted this as evidence of the non-utility

[13] P. M. Symonds, The measurement of teaching efficiency in high school, *Ed. Adm. & Sup.*, 1927, *13*, 217–231.

[14] Boyce, A. C., Methods for measuring teachers' efficiency, *14th Yearbook, Natl. Soc. Study Ed.*, Part II, 1915.

[15] L. M. Crabbs, *Measuring efficiency in supervision and teaching*, T. C. Cont. Ed., No. 175, 1925.

of ratings, but Symonds convincingly shows that it really exhibits the inadequacy of the AR for this purpose. His reasons are twofold: (1) The AR has a low reliability (approaching zero) as compared with the reliability of standardized tests themselves. This is easy to demonstrate statistically and is based largely upon the presence of a cumulative error. (2) "A second criticism which concerns the validity of the AR as a measure of teaching efficiency is the degree to which a study correlates with intelligence. The AR is used as a measure of teaching efficiency so that correction can be made for pupils or classes of different mental ability. If, however, there is little relationship between specific ability in a subject and general mental ability (as appears to be the case in drawing, manual arts, etc.), all validity is lost."

Thus ended one of the most promising ventures in educational science. Discouraged, many persons inquired, "If the AR cannot be trusted to solve the difficulty, what else can?" Temporarily baffled, it seemed necessary to retreat, re-examine one's premises and attack anew with different weapons. Courtis's[16] interesting application of the Gompertz growth curve to scholastic progress has opened up new possibilities, but until it is confirmed and better understood as a general law of biological development, one would do well to be cautious in its use. By means of a rotation arrangement, Courtis measured two very good and two very poor cadet teachers for ability to produce student growth in spelling after equal practice. A comparison of the plotted slopes showed that the rate of growth under the best teacher was twice that under the poorest. The method certainly deserves to be refined and extended further.

The most exhaustive recent study conducted along the lines established in this preliminary analysis as requisite for sound experimentation in this field is an impressive monograph by Taylor.[17] A little extract from his historical sketch shows how the insistence upon greater objectivity and detachment has steadily grown more pronounced during the last two decades. "Davidson, in 1913, before the National Educational Association, advocated judging teachers by the effects produced. Kent, in 1920, protested against the small place given to results of instruction in all rating schemes.

---

[16] S. A. Courtis, The measurement of the effect of teaching, *School and Society*, 1928, *28*, 52–56; 84–88.

[17] H. R. Taylor, Teacher influence on class achievement: A study of the relationship of estimated teaching ability to pupil achievement in reading and arithmetic. *Genet. Psychol. Mon.*, 1930, VII, No. 2.

He argued that the product of teaching could be measured objectively and that such evaluation should constitute the major portion of any plan for teacher rating. Courtis, in 1921, proposed that teaching ability be defined wholly in terms of the changes to be produced in children. Monroe and Clark, in 1921, in their bulletin on 'Measuring teaching efficiency,' say, 'An ideal plan would measure only the modifications produced in the pupils by the teaching process.' They state that all elements of pupil growth—ideals, interests, attitudes—must be considered, as well as skill and knowledge.'' (Page 92.)

It is interesting to observe the progress of investigative technique which has taken place within the last decade by comparing Taylor's dissertation with an article by Hill.[18] The latter author made an extensive study of several city school systems in which he checked ratings for 135 teachers against standard tests in arithmetic, penmanship, and spelling. The correlation between the ratings and class results was .454. Instead of lamenting this low figure, one can only marvel that it was obtained in view of the fact that differences in ability of the classes were not equated and that the rating indices were seriously attenuated by being lumped into but three categories!

Taylor's own research attempted to answer a similar question: Are such estimates of teaching ability as are commonly available in city school systems indirectly measures of the relative accomplishment of the pupils? He secured ratings of 133 teachers of grades 4–B to 8–A made by five superiors. Since the children in these classes totalled 1968 a large number of comparable groups could be constructed. Standardized tests in the two disciplines under investigation were administered at the beginning and end of a semester, but despite the employment of the most refined statistical procedures of the Stanford school, only modest and low relationships between the significant variables emerged. It would be unkind to cite the old saw of the mountains in labor, but the author's simple conclusions deserve quotations: ''The assumption that estimates of teaching ability are measures of the merit of a teacher because they are indirectly measures of the proficiency of pupils traceable to differences in teacher influence on achievement is to a considerable extent justified with reference to reading achievement but only slightly with reference to achievement in

[18] C. H. Hill, Efficiency ratings of teachers, *Elem. School J.*, 1921, *21*, 438–443.

arithmetical computation. The better teachers in this field are, on the whole, inadequately recognized. This finding is the major contribution of the investigation." (Page 163.) Perhaps it will not be amiss to call attention to the fact that whatever slight validity ratings possessed was unearthed by matching them against achievement test results—and not *vice versa*.

It must not be assumed from the content of the preceding discussion that the measurement of teaching efficiency is confined to the dilemma represented by rating scales at one horn and achievement tests at the other. There are still a few other sources of estimates which have been used or proposed with indifferent success, such as the number of pupils who fail a given teacher's course or the number who pass certain major trials of educational status such as the Regents' and College Entrance Board Examinations. Although there are still benighted places where an instructor's competence is determined by this absurdly pragmatic test, it is utterly lacking in scientific "respectability" if institutional standards and degrees of native capacity are not equated for all. A far more sober proposal is that which suggests the percent of students electing to continue the next higher course as an index. For instance, suppose three instructors, A, B, and C, each teach a section of introductory philosophy to groups of equal talent: if 10% of A's class voluntarily continues with another offering in the department of philosophy as compared with 5% for B and 2% for C, the order of merit seems clear, at least with respect to inspirational powers. Indirectly, then, a teacher is measurable by the amount and quality of patronage and discipleship he arouses, just as Socrates was glorified by Plato and Plato in turn by Aristotle. The reputed merits of Wundt, Titchener, and Cattell, to take but a few recent instances, have been in part built upon the competence and future eminence of the Ph.D.'s they helped to train. Unfortunately, this criterion is complicated by institutional and social factors which make more for uniqueness than comparability and it has the special disadvantage of being applicable only to the upper levels of the educational ladder. This index, then, while capable of development under appropriate conditions, does not appear as immediately promising as a broadly-conceived testing program.

It is to a description of our special experimental situation and the all-important battery of measuring instruments that we must now turn.

CHAPTER III

THE EXPERIMENTAL PROCEDURE

The general plan of the investigation about to be described conforms to that commonly known as the parallel-group experiment. The following college circumstances made such an attack practicable: At the beginning of the second semester of the academic year, three sections of a required sophomore course in educational psychology were scheduled to be taught by three instructors whom we shall designate hereafter as A, B and C. All the professors involved were experienced men, holders of the Ph.D. degree, who had taught this particular course a number of times in the past. Since they were attached to the School of Education of a representative state institution we may presume that these three individuals possessed at least average university scholarship and pedagogical skill.

The major purpose of this study being the measurement of the relative teaching efficiency of A, B and C by means of data indicating the amount of desirable personality growth which had taken place in the students in their classes, great care had to be exercised to see that the same essential conditions (exclusive of "method") obtained in all groups. All sections by agreement used the same text-book and supplementary reference shelf, and although the precise classroom procedure was left free to be determined by each professor, the lecture-discussion method was the principal agency of instruction followed by all. The records of C, the writer of this report, must be considered somewhat apart from those of A and B, since he was the person responsible for this experimental set-up and the only one fully aware of the total situation. His colleagues, A and B, who may be unreservedly compared with each other, were induced to cooperate in the extensive testing program demanded, on the plausible pretext—which, as a matter of fact, was a significant by-product of this activity—that the character of the shifts in students' attitudes as a result of a college career was the main objective sought.

C's class contained 105 students, just about twice as large as A's and B's (A = 55 and B = 59); men and women were equally represented in all instances. This was deliberately arranged in order to throw light on the "quantity" factor of the efficiency formula by consolidating two ordinary-sized sections and using the three hours of the instructor's time thus saved for ten-

minute interviews with each student. This allowed at least two personal conferences during the term: in the first, a program of specialized reading or an individual experiment was outlined after consultation to determine the student's interests, and in the second, his "project" was heard and discussed. C concerned himself primarily with developing the student's level of scholarship in the subject of instruction and made no direct attempt to influence the attitude of his charges, for this would have introduced a disturbing element in the measurements applied. This may appear to make his records incomparable with the others, but since it is one of the premises of this research that the efficiency of a teacher cannot be divorced from the efficiency of the method he spontaneously elects, whatever superiority of one group to any other is discovered would be as much attributable to the instructor's efforts as to the technique employed. In all higher institutions of learning, the instructor is normally free to choose whatever "method" he considers best—in fact, his own rating is in part dependent upon the wisdom he displays in selecting the most appropriate teaching procedure. The method, in other words, is contingent upon the teacher, and not the converse.

On the principle that concomitant or incidental learning is as important as any immediate or direct mastery of content in judging the social and individual value of a course—a view which is much more attractively expressed by Principal Jacks' insistence on the "education of the whole man"—the battery of tests to be applied must be sufficiently comprehensive to measure these "secondary" as well as the "primary" outcomes of instruction. If it be correct to assume that the efficient teacher is one who can not only influence development in a good direction in a segment of the pupil's nature, but who can also affect beneficially the life pattern of the *total* personality, then more than a plain achievement test is needed. The additions required must yield evidence of differential growth in social and emotional equipment and in such quasi-intellectual factors as attitudes, interests, and value hierarchies.

After consideration of the theoretical requirements of the problem, the availability of suitable and reliable tests, and the expediency of securing responses from several hundred undergraduates, the following measuring instruments were chosen:

(1) The Allport-Vernon Scale of Values Test.
(2) The Strong Vocational Interest Blank.
(3) The Bernreuter Personality Inventory.

(4) The Watson Test of Public Opinion (abridged).
(5) An Achievement Test in Educational Psychology.

The hypotheses upon which the selection of each test was based are as follows:

(1) The Allport-Vernon scale of values blank is designed to differentiate the six fundamental types of personality as outlined by the eminent contemporary German philosopher and educational theorist, Eduard Spranger, in his volume, *Lebensformen.* According to Spranger, the best clue to the nature of any personality is obtained from a knowledge of the things which he prizes most highly. In his opinion, this criterion permits us to distinguish six, and only six, main personality patterns: The *theoretical,* or the man who loves truth and facts beyond all else; the *aesthetic,* to whom beauty and the appearance of things is most significant; the *political,* who finds his highest joy in the exercise of control over others; the *social,* who considers friendship and human relations with others most precious; the *economic* individual, who esteems practicality above all other goods; and the *religious* man, who seeks his chief delight in contact with the ruling power of the universe. The Allport-Vernon blank has cast this intuitive grouping into a form whereby the subject checks his preferences among several possibilities of action. A profile for each person, which is readily obtained by plotting the frequency of choices of a certain character against the "norms" for the general population, shows purely the relative scale of worth of objects or activities to a single person, and gives no hint, unfortunately, of the affective intensity with which *all* values may be held. However, the serviceability of this blank for the present study is clear when we envisage the possible results of learning the content, technique, and spirit of educational psychology. As customarily taught, educational psychology professes to be a scientific discipline and if more than superficially assimilated should contribute something to the formation of the *theoretical* view-point. That teacher, then, who raises the final level of theoretical-mindedness among his students over what it was when the course began may be considered more effective, other things being equal, than one who leaves them with the same standpoint with which they began.

To be sure, the nature of the valuational change depends upon the character of the course pursued. In the absence of any unambiguous evidence on this point, it was considered wise to see whether

the *aesthetic* value would become relatively more pronounced after a course in English *literature;* the *political,* after a course in *political science;* the *social,* after a term's work in *sociology;* the *economic,* after a semester of *applied* psychology; and the *religious,* after a period of voluntary Bible class study. The data were obtained in an *unwissentlich* manner with the help of instructors in just such courses by securing student responses to the scale during the first and last weeks of the semester. With the exception of the religious group, which was too small and irregular in attendance to be helpful, interesting material was secured from all these classes and will be described in Chapter V.

Some may object that this application of the Allport-Vernon test makes no provision for a control group to rule out the effect of mere maturity. While it is doubtful if chronological age as such can ever act as a causative factor and even more doubtful whether any two environments in our complicated modern life can be equalized, still some light can be thrown on this question by observing changes in value scores for a non-college group of equal age and native ability over the same period of time. Through the aid of the personnel department of the United States Aluminum Company's plant at New Kensington, Pa., twenty-five apprentices in a foreman-training class were measured in this way. These lads were all high-school graduates with Otis scores comparable to the average college sophomore, who for temperamental or financial reasons had entered industry directly rather than by the route of the engineering school. As shown by Table I below, there is noth-

TABLE I

SCORES MADE ON THE ALLPORT-VERNON SCALE OF VALUES TEST BY 25 FACTORY APPRENTICES IN FEBRUARY AND MAY, 1933

	Theoretical	*Economic*	*Aesthetic*	*Social*	*Political*	*Religious*
Initial Mean Scores	33.22	33.74	22.26	28.87	31.17	30.87
S.D.	7.50	7.59	6.66	5.10	5.51	7.60
Final Mean Scores	32.96	33.52	22.39	29.04	29.48	32.61
S.D.	7.31	4.77	6.41	6.71	5.04	7.43
Amount of change	–.26	–.22	.13	.17	–1.69	1.74

ing in the character of the average scores made in February and May which gives the least hint that any marked shift in their standards of value had occurred.

Since 30 constitutes the "norm" on each value, it is clear that this apprentice group is quite representative in its preferences with the exception of the low average *aesthetic* scores, the extraordinary stability of which testifies to the diagnostic character of the test, since it is in just this respect that one would expect a factory population to rank lowest. On the basis of this sample, it seems reasonable to suppose that any changes in average *aesthetic* level occurring during a course in English poetry may be directly attributable to influences emanating from it and not from other collegiate or extra-mural stimuli; and *pari passu,* any similar alterations in amount of theoretical-mindedness consequent upon a term's work in psychology may be assumed to have an obvious and definite origin. It is curious that this problem is never raised with achievement tests although it is possible for one's information as well as one's attitudes to be affected by "outside" sources. In both cases, we assume that where paired groups are involved, these external opportunities for learning are equal.

(2) The Strong Vocational Interest Blank was chosen on the assumption that a course in educational psychology, replete as it is with both written and oral references to teaching situations, should make the students as a whole more pedagogically-minded and increase their interest in the problems and activities of such a career. The Strong test, in this case, was administered not with any direct occupational or educational guidance in mind, but to see how much more closely the students approximated in their likes and dislikes those which are known to characterize the more distinguished members of the teaching profession. Presumably, a partial measure of an instructor's effectiveness may be obtained from the degree to which his pupils progress toward that goal. Of course, a student may enter a given class with an inadequate knowledge of its content and later discover that he is not adapted to the career for which it prepares. A retrogression in such a student's case cannot be interpreted as a reflection upon the instructor. However, our *mass* use of the instrument seems justified, for certainly if a medical course made the *average* member enrolled less interested in the work of a physician, or the average art course diminished one's love for the beautiful, it is to that extent a distinct human failure. Stated in the negative form, the acceptability of this postulate is

unquestioned; consequently, the most efficient instructor in this particular course will do most to deepen and confirm a student's embryonic interest which makes all beginning possible.

(3) The third major item in the test collection was the Bernreuter Personality Inventory. This new instrument has the high technical advantage of brevity and convenience. From the nature of the answers, a respondent's standing in four different traits may be simultaneously determined, viz., neurotic tendency, introversion-extraversion, degree of self-sufficiency, and dominance-submission. For the purposes of this investigation, the blanks were scored for neurotic tendency, since that was the only one of the four traits in which an alteration—in this case, a diminution—could be considered ethically, socially, or biologically desirable. It is the only feature of the composite which represents a distinct personality liability. If the mental hygiene movement has any educational significance at all, it implies that unfortunate emotional habits among the school population need to be supplanted by saner modes of reaction if the danger of serious psychiatric disorders are to be circumvented. Since a course in educational psychology bears some reference to this topic, it seems just to say that the most efficient teacher will bring the largest percentage of his students to the level represented by a balanced, stable, and well-adjusted personality; or, if this seems too fantastic, he will at least reduce thoroughly the number of traumatic symptoms exhibited.

(4) The Watson test of public opinion, as every one among the initiated realizes, is essentially a measure of prejudice. It appeared desirable to secure some knowledge of the changes occurring in this respect since most educational philosophers agree that a reduction in the amount and kind of irrational political, economic, social, and religious prejudice is a desideratum of all good schooling. Those pupils who make the largest mean gain in tolerance presumably have the best instructor. Watson's original blank was a bit too long for the exigencies of the testing situation, hence, a modification comprising the generalization part, the degree of truth part, and the "cross-out" section was used instead. The inference, moral judgments, and arguments sections being omitted meant that the revised form was only about one-half the length of the original. While this abbreviation is probably gained at the sacrifice of a certain amount of validity and reliability, this loss should not be exaggerated, since it is known that many of the shorter forms of the Army Alpha Intelligence Examination have approximately the same factors of advantage as the unabridged edition.

(5) An achievement test in the subject-matter of educational psychology constitutes an indispensable part of any exhaustive survey such as has been undertaken here. Long the only outcome which could be determined precisely, it is still a peer and not a subordinate of the other indices of efficiency. Here, accurate scholarship—surely a worthy aim of all instruction!—has an opportunity to reveal itself. If all our measures are to be attitudinal in character, there is grave danger that the inspirational but "unsubstantial" type of teacher may possess an undue advantage over his less spectacular but solider colleagues. Hence, the objective test which had been used for a number of years at this institution as a part of the final examination and which had profited from the joint criticism of the staff members in its final arrangement was introduced. This consists of 150 true-false statements, has a known reliability of .92, and a validity (as determined by the correlation with semester grades in the subject) of .78. One need hardly mention that the students under the tutelage of the most capable teacher will in some way have acquired the greatest amount of facts, the richest body of information, and the largest number of relevant associations.

On the basis of pupil gains as measured by the differences between properly matched mean initial and final scores in these five tests, the several instructors involved may be ranked for teaching efficiency. In the present state of scientific development, there does not appear to be any more objective or exact way of measuring instructional ability. A grave limitation is the uncertain weight to be attached to each of these five changes. Shall they be pooled as equivalent and each given an arbitrary weight of one, or do these five types of alteration differ among themselves in importance and desirability? Is a gain of .5 S.D. in teaching interest as significant as a similar advance in psychological knowledge? Does a ten-point loss in neurotic tendency mean as much as an equal loss in prejudice? This is an intricate but decidedly fundamental issue and one which will probably not be answered until the advent of experimental axiology, if such a science be possible. It is a question which even the wise compilers of the Seven Cardinal Principles of Secondary Education failed to raise, much less settle. Whether character is to be preferred to health is a dilemma which must be solved individually in the absence of a footrule or a higher common magnitude to which both may be referred.

But the nature of our problem forbids any evasion of this difficult matter. Serious injustice may be done if we say that a given

teacher is more efficient than another because his pupils improved noticeably in three of our five tests while his colleague's made equal gains in only two! Perhaps those two were more "vital" and occupied a higher place in the hierarchy of meaning than the remaining three. There is, unfortunately, no way of determining this save by recourse to expert judgment, to which one is inevitably and reluctantly driven. Since some rating of the relative importance of these five changes was required, twelve judges, all professors of education or psychology whose names appear in *Leaders of Education,* were requested to rank the outcomes of a professional course in educational psychology according to the order of merit method. Each judge was given a slip containing the phrases below and told to mark the most important outcome 1 and the least significant 5:

Broad love of truth.
Interest in teaching.
Well-regulated emotions.
Freedom from social prejudices.
Accurate knowledge of educational psychology.

The final placement of these goals will obviously yield a clue as to the relative weighting of the corresponding tests, which was secured by Hull's method for transmuting scores as described in H. E. Garrett's *Statistics in Psychology and Education,* pages 111–115. Table II is derived from the results of the combined judgments.

TABLE II

RANK ORDER OF MERIT FOR FIVE OUTCOMES OF A PROFESSIONAL COURSE IN EDUCATIONAL PSYCHOLOGY AS DETERMINED BY TWELVE JUDGES

Objective	*Mean*	*P.E.*	*Rank*	*Weights*
Truth	3.08	.93	3	140
Interests	3.00	.57	2	140
Emotions	4.25	.77	5	100
Prejudices	3.33	.61	4	129
Information	1.33	.46	1	203

The reader will understand that all five tests in this series were administered twice in identical forms—once in all three classes during the third week of February, 1932, after the enrollment for the respective divisions had been stabilized, and again during the last

week of May, 1932.[1] The achievement test was given as part of the final examination in order to ensure maximum motivation but its results were not used in calculating grades. A period of three months seems all too brief for accomplishing the attitudinal changes proposed and some would claim that even the span between freshman and senior years is hardly long enough to permit alterations of any magnitude to occur. However, education must be recognized as to some extent a forcing process. In an institution like a college, the human plant is in an intellectual hothouse which cannot be matched elsewhere. Moreover, if our measuring instruments are assumed to be sensitive enough to detect changes over a four-year period, they should be able to reveal the presence of any shifts within one-eighth of that interval.

The necessary matching of individuals in all classes on the basis of the pre-test scores was made separately for each of the five principal measures. It was desired to make the equating as closely as possible without sacrificing cases and the effort was fortunately quite successful since the differences between all the pairs fell well within the limits of the probable error of measurement of each test score as the figures in the appropriate tables to follow indicate. Matched groups were thus set up for A versus B, A versus C, and B versus C; the pairs varied in number from 27 to 38. This would have been difficult to do but for the fact that C's group was about equal in size to A's and B's combined. In all multiple matching such as here attempted, overlapping inevitably occurs, *i.e.*, the A individuals who were matched with B for each trait, were only in part represented again in the A group when matched with C. This happens because an individual in group A with a score of 95 can be paired satisfactorily enough with a B person with a score of 94, while an A individual with a score of 70 may not find a mate in B but can often locate an approximate partner in C.

From this stage on, the enterprise was chiefly clerical and does not merit detailed description, since the statistical constants and the tabular captions will generally indicate the nature of the calculations involved. Obviously, the center of interest lies in a comparison of the initial and final means in all tests.

---

[1] The scoring of the papers was a tremendous task and would never have been accomplished but for the competent and interested help of several assistants. Over two hundred individuals took five tests twice, making a total of more than two thousand blanks to be marked and checked, a process which was especially laborious in the case of the Strong inventory, although the Allport-Vernon and Bernreuter booklets, because of their peculiar weighting techniques, were almost as time-consuming.

CHAPTER IV

RESULTS AND DISCUSSION

For purposes of preliminary comparison, the first definite data pertaining to the relative effectiveness of the three instructors were obtained by correlating the raw intelligence scores of the students with their standing in the final achievement test. This was done on the assumption that the class which showed the highest correlation between these two measures was taught by an instructor who was eliciting the maximum performance from each individual. A similar hypothesis has been employed in measuring the effect of homogeneous grouping upon accomplishment with the usual finding that the average correlation for sectioned classes is appreciably higher than with heterogeneous groups. If this be a sound way of assessing an administrative device, there is no reason why the same test of merit cannot be applied to the measurement of instructional skill. Where the correlation between mental level and accomplishment is relatively high, the teaching must have called forth the best efforts of most students, regardless of variations in capacity; if this relationship is absent, a considerable amount of "loafing" or irregularity in personal treatment must have occurred. With this in mind, Table III should serve as an exploratory test of variations in teaching capacity. While the probable errors of these coefficients are large, the chances that B and C at least would reverse places are negligible. It is conceivable that these magnitudes reveal only the "hard driver" and the lenient instructor in the trio, but if work in accordance with one's capacities is a desideratum, then the "hard driver" has at least this pedagogical sanction on his side.

TABLE III

CORRELATION BETWEEN INTELLIGENCE AND ACHIEVEMENT TEST SCORES IN THREE CLASSES OF EDUCATIONAL PSYCHOLOGY

Instructor	*r*	*P.E.*	*N*
A	.32	.11	38
B	.10	.11	37
C	.44	.06	81

A wholly different picture of the relative merits of these three instructors resulted when in place of the objective check used above

student ratings of the teachers were gathered. A departmental assistant located ten students who had taken work with all the instructors involved in this analysis as well as with two others of the Liberal Arts faculty (D and E). These students filled in the Purdue Rating Blank for college instructors anonymously, but with directions to be as fair as circumstances and their knowledge of each professor permitted. While the estimates of any one rater are to be heavily discounted, the average judgment of a number of students independently given is worthy of serious consideration. Table IV presents the results of this inquiry.

At first glance one is struck by the fact that the rank order of merit in Table IV is exactly the reverse of Table III. Does this imply that the students attach a "halo" to the "easy" instructor, assuming that the findings of Table III justify one in considering B such? Perhaps; but more likely this discrepancy among the indices of effectiveness suggests the absolute necessity of accumulating as much evidence as possible about each teacher to be rated. Despite all we know about the positive correlation of good abilities, it is perfectly possible for an instructor to rank high in one teaching attribute and low in another. Moreover, the traits are undoubt-

TABLE IV

AVERAGE ESTIMATES OF THE TEACHING TRAITS OF FIVE INSTRUCTORS PREPARED BY TEN UNDERGRADUATES USING THE PURDUE RATING BLANK

Instructor:	*A*	*B*	*C*	*D*	*E*
Trait					
1. Interest in subject	94	95	96	93	92
2. Sympathetic attitude toward students	75	90	71	58	91
3. Fairness in grading	73	96	83	65	84
4. Liberal and progressive attitude	87	85	89	86	83
5. Presentation of subject-matter	84	93	70	92	77
6. Sense of proportion and humor	82	90	68	87	81
7. Self-reliance and confidence	92	91	93	94	84
8. Personal peculiarities	87	86	76	76	89
9. Personal appearance	93	84	74	69	78
10. Stimulating intellectual curiosity	80	85	69	85	68
Grand Mean	84.7	89.5	78.9	80.5	82.7

edly of different degrees of cruciality, so that the legitimacy of the grand means in Table IV is a bit questionable. The averages for each trait, however, do not suffer from this limitation.

The high specificity of teaching talent is suggested by examining the outstanding relative excellences of A, B, and C, respectively. A leads in two features (items 8 and 9), B in five (items 2, 3, 5, 6, 10), and C in three (traits 1, 4, 7). These peaks and valleys in the total individual profile argue against the view that instructional skill may be a "g" factor in Spearman's sense—a conclusion which we shall find supported, if not confirmed, by the nature of the objective test results. We must plainly avoid condemning or lauding a man on one count alone.

The limitations of student or other types of rating stand out sharply when confirming evidence of some of the judgments is sought through different channels. For example, item 3 in Table IV (dealing with the matter of "fairness of grading") can be checked to some extent by noting how closely each instructor's final grades in the course correspond with the scores made by his pupils in the achievement test. While it is true that other considerations may on occasion legitimately affect the marking and that the accomplishment test is an adequate measure of factual information only, the virtues of an independent criterion of performance should not be minimized. The pertinent data appear in Table V.

TABLE V

CORRELATIONS BETWEEN FINAL MARKS GIVEN BY DIFFERENT INSTRUCTORS IN EDUCATIONAL PSYCHOLOGY AND SCORES ON AN ACHIEVEMENT TEST

Instructor	*r*	*P.E.*	*N*
A	.08	.09	52
B	.24	.09	47
C	.90	.01	103

Even if one views the accomplishment test as no better than a good final examination, past studies would lead one to expect a closer relationship between rank in a final examination and average position in daily work or earlier quizzes. The value for C is the only one which approaches this expectation. Unless this reasoning is spurious, B can hardly be the "fair marker" he is assumed to be.

However, the three preceding tables yield only the most tentative and uncertain conclusions concerning the different teachers,

and the need for supplementary evidence becomes all the more acute as a result of this preliminary exploration. Having acquired a perspective upon the problem, we shall be better prepared to appreciate the significance of the major findings of this study. In order to bring the data together into as compact a form as possible for the sake of facilitating comparisons, Table VI has been constructed. In it will be found the gist of the present investigation. The reliabilities of the differences between the means at the beginning or end of the semester (or gains) constitute, of course, the heart of the matter. The final average values are greater than the initial in all instances except for *prejudice* and *neurotic tendency* where a "loss" is obviously a "gain." Section 6 of the table is an addendum to section 5, since it seemed desirable to match students not only according to initial knowledge of psychology, but also according to native ability, both bases for equating being applied separately.

It is easy to lose sight of one's main objective in the presence of such an assemblage of figures, and it is therefore important that the table be analyzed in detail for each instructor and each measured outcome. The rigorous statistician will first be impressed by the failure of every one of the differences between the means to conform to the conventional requirements of "certainty," although some closely approach them. The psychologist cannot fail to notice the fluctuations in competence with each variable. No one among the three instructors excels the other two in *all* respects and every one leads in at least one item. For the practically-minded administrator, this state of affairs is most conveniently portrayed by Table VII (*infra*), which is simply a summary of the comparative excellences revealed by each instructor in the various traits; *e.g.,* in the first measure (Allport-Vernon test) teacher B excelled *both* his mates as indicated by the number 2, C excelled one of his companions, and A none. The totals at the bottom of the columns indicate the number of times each man emerged a "winner," regardless of whether this victory was by a close or a wide margin.

The long conventional formula for computing the standard deviation of the obtained difference between the means of any two groups has been used throughout Table VI. An informed critic may object that *greater* reliability would have been found for the differences had the S.D. of the Mean *Gain* and the S.D. of the difference between the Mean Gains been computed. This is quite probable, since there is enough correlation between the students'

TABLE VI

SEMESTER CHANGES IN GROUP MEANS OF CLASSES TAUGHT BY THREE DIFFERENT INSTRUCTORS WHEN PAIRED INDIVIDUALLY FOR INITIAL STANDING IN SIX EDUCATIONALLY SIGNIFICANT FUNCTIONS (NOTE: COLUMN L STANDS FOR "MATCHING LIMITS")

Pair	*N*	*L*	*Initial Mean*	*S.D.*	*Final Mean*	*S.D.*	*Mean Gain*	*Difference in final averages*	$\sigma_{av.F}$	$\sigma_{diff.}$	$D/\sigma_{diff.}$	*Changes in 100 against reversal*
I. VALUES ($r_{IF} = .66$)												
C	34	1	27.44	6.83	28.68	7.69	1.24		1.32	1.01	.20	58
with B			27.10	6.53	28.88	6.91	1.78	.20	1.19			
C	39	1	28.22	5.74	28.86	5.90	.64	.35	.94	.797	.44	67
with A			28.26	5.76	28.51	6.54	.25		1.04			
B	31	1	28.35	6.28	29.84	6.91	1.49	.71	1.24	.965	.74	77
with A			28.41	6.24	29.13	6.52	.72		1.17			
II. INTEREST ($r_{IF} = .70$)												
C	32	12	138.22	81.85	161.06	112.35	22.84	8.44	19.86	14.49	.63	74
with B			138.57	81.77	152.62	95.33	14.05		16.85			
C	36	12	95.31	90.96	139.89	111.94	44.58	10.39	18.66	13.36	.78	79
with A			95.61	89.36	129.50	84.46	33.89		14.08			
B	29	17	119.10	65.97	120.72	83.02	1.62		15.42	11.24	1.34	91
with A			117.92	71.44	135.80	79.18	17.88	15.08	14.70			
III. NEUROTICISM ($r_{IF} = .81$)												
C	36	9	−47.57	49.28	−66.20	55.56	−18.63	5.11	9.26	6.52	.78	79
with B			−48.91	51.11	−71.31	66.65	−22.40		11.11			
C	36	9	−30.42	53.82	−38.95	45.85	−8.53	7.83	7.64	6.15	1.27	89
with A			−30.47	51.49	−46.78	53.89	−16.31		8.98			
B	27	14	−23.30	48.50	−43.17	57.30	−20.87	8.83	11.03	6.47	1.36	91
with A			−22.69	49.74	−52.00	46.79	−29.31		8.99			
IV. PREJUDICE ($r_{IF} = .78$)												
C	31	7	84.47	27.46	79.67	25.79	4.80	6.54	4.61	3.4	1.92	97
with B			83.81	28.27	73.13	24.84	10.68		4.46			
C	33	7	78.1	31.38	66.4	33.07	11.7		5.76	4.27	.43	66
with A			78.0	29.83	68.24	32.52	9.76	1.84	5.66			
B	28	7	88.0	34.78	75.68	30.61	12.32		5.78	5.02	.64	74
with A			88.61	33.41	78.89	38.01	9.71	3.21	7.18			

TABLE VI.—(*Continued*)

Pair	*N*	*L*	*Initial Mean*	*S.D.*	*Final Mean*	*S.D.*	*Mean Gain*	*Difference in final averages*	$\sigma_{av.F}$	$\sigma_{diff.}$	D/$\sigma_{diff.}$	*Changes in 100 against reversal*
V. Educational Psychology (r_{IF} = .52) (equated initially for knowledge of educational psychology)												
C			75.65	10.40	95.25	11.49	19.60		1.94			
with	35	4								1.74	2.39	99.2
B			75.08	9.91	99.40	9.12	24.32	4.15	1.54			
C			64.72	12.59	93.44	12.52	28.72	1.17	2.29			
with	30	4								2.23	.52	70
A			64.34	12.74	92.27	12.37	27.93		2.26			
B			65.76	10.46	95.93	9.94	30.17	3.37	1.78			
with	31	3								1.784	1.89	97
A			65.76	10.20	92.56	10.38	26.80		1.86			
VI. Educational Psychology ($r_{int.-ach.}$ = .33) (equated initially for intelligence)												
C			98.26*	19.75	98.34†	11.53	.....	3.03	1.87			
with	38	5								2.00	1.52	93
B			98.69	24.16	95.31	9.61			1.56			
C			95.82	23.46	90.16	10.14	.....	.79	1.76			
with	33	5								2.049	.27	60
A			95.42	19.46	89.37	10.25			1.78			
B			96.88	21.21	94.25	9.65	.....	4.68	1.66			
with	34	5								2.304	2.03	98
A			98.18	21.74	89.57	12.12			2.08			

* This column shows the intelligence test scores.
† This column lists the final group means in educational psychology.

initial and final scores to make it appear that the S.D. of the Gains would be relatively small. Also, the procedure of using actual gains would seem somewhat fairer than that of disregarding the slight original differences in the Initial Means in the matched groups and using the Final Means without correction.

This possibility was tested by repeating the computations for section V of Table VI. The method used, however, while not exactly the same as that outlined in the preceding paragraph, is equivalent to it in effect. The formula

$$\sigma_{d_{mg}} = \frac{\sigma_{dg}}{\sqrt{n}}$$

indicates that the S.D. of the difference between the mean gains of any paired groups is equal to the S.D. of the matched gains

($dg = g_1 - g_2$) divided by the square root of the number of pairs. One important advantage of this method is that it takes full account of the correlation between the first and second administration of the two measures without the necessity of using the r's in the computation. Table VIa shows the results obtained from the same raw data by the use of this formula:

TABLE VIa

SECTION V OF TABLE VI WITH THE ESSENTIAL COMPUTATIONS REPEATED, GIVING DIFFERENCES IN GAINS MADE IN KNOWLEDGE OF EDUCATIONAL PSYCHOLOGY UNDER DIFFERENT INSTRUCTORS. NUMBER OF CASES, MATCHING LIMITS, AND PAIRINGS ARE THE SAME IN BOTH INSTANCES

Groups	*Mean Gains*	*Difference*	$\sigma_{d_{mg}}$	$\frac{Diff.}{\sigma_{d_{mg}}}$	*Chances*
C	19.36				
with		4.61	1.93	2.39	99.2
B	23.97				
C	29.89				
with		1.50	2.50	.60	73
A	28.39				
B	30.00				
with		3.31	2.87	1.15	87
A	26.69				

A glance at the "chances" column shows that the probabilities that true differences between the groups are really present are not noticeably altered by an alternative method of computation: the "critical ratio" in one pair is unchanged, in another slightly raised, and in a third, definitely lowered. The results obtained with either method do not diverge sufficiently to warrant any preference for one or the other set of values. It is somewhat disappointing to be denied the satisfaction of unearthing strictly reliable differences, although the chances against a shift in the direction of the obtained superiorities vary from 58 in 100 as the lowest to 99.2 in 100 as the highest. There is some reason for believing that the requirement of three times the standard error is needlessly rigid and severe; as a basis for practical decisions, adoption of the arbitrary psychophysical "limen" at 75 in 100 has much to recommend it, for it permits action in accordance with the best information available. There are many situations in life where we must be content with probabilities.

The method of comparing the variations in teaching efficiency used in Table VII, while easily comprehended, suffers from a serious defect—it credits an instructor with a unit of superiority

whether his final mean exceeds his competitor's by only a fraction of a standard deviation or by a substantial multiplier; *e.g.,* in item 5 of Table VI, the S.D. of the difference between A and C is only .52 but between B and C it is 2.39. Clearly, some means of weighting the "degree" of superiority according to the amount of preponderance is required. The degree of excellence is crudely, but perhaps most conveniently, expressed by the size of the chances that the obtained magnitude of difference between any two teachers would not be reversed. One teacher may have 66 chances in 100 of being better than a second in a certain feature and 74 chances in 100 of being better than a third participant;—yet can this properly be considered as evidence of greater effectiveness than *one* superiority with 97 chances in 100? Judgment enters heavily here. Most people would probably decide that it is symptomatic of greater skill to win two closely-fought contests than to emerge overwhelmingly victor in but one. How much better, however, is hard to determine. The method of "chances" at any rate obviously gives definitely greater weight to two mild superiorities than to a single outstanding one (since 66 plus 74 exceeds 97).

TABLE VII

RAW OR GROSS SUPERIORITIES OF DIFFERENT INSTRUCTORS IN THE ABILITY TO PRODUCE CERTAIN DESIRABLE CHANGES IN STUDENTS OF THE SAME LEVEL OF ATTAINMENT AT THE BEGINNING OF A COURSE
(SUMMARIZES TABLE VI)

Teacher:	*A*	*B*	*C*
Pupil scores paired in			
(1) Allport-Vernon Theoretical Value	0	2	1
(2) Strong Interest Blank	0	2	1
(3) Bernreuter Personality Inventory	0	2	1
(4) Watson Fair-Mindedness Test	1	0	2
(5) Achievement in Educational Psychology	2	1	0
(6) DeCamp Intelligence Test	0	1	2
Total of superiorities	3	8	7

Even this gives only an incomplete picture of the relative excellence of each instructor since the objectives (according to Table II) are not of equal merit. Consequently, the degree of superiority as measured by the probabilities against reversal must be multiplied by the comparative weights of the separate outcomes. These computations were performed and organized in Table VIII. Instruc-

tor C's composite score, *e.g.*, in prejudice (8514) was obtained by multiplying 66 (the "chances")[1] by 129 (the judges' mean estimate of importance). The sum of these composite scores should yield some indication of the "general" teaching efficiency of the three men. The numbers reveal, of course, only the relative standing of the teachers compared and by no means imply that B is more than three times as good a teacher as A.

It will be observed that the weighted value of "knowledge of psychology" is apparently doubled by including both sections (5) and (6), but since it is impossible to average these two factors without doing the same to other cases involving dual superiorities the only way to correct for this is to compute the totals once by omitting (6) and again by omitting (5). A reduction of A's relative inferiority occurs in either case; omitting (5) gives C a slight margin over B, but omitting (6) gives B a larger margin over C. No matter by what method the data are treated, B generally emerges first, followed closely by C, with A definitely last.

TABLE VIII

RELATIVE TEACHING SUPERIORITY OF THREE INSTRUCTORS WHEN WEIGHTED BY THE CHANCES AGAINST REVERSAL OF COMPARED MEANS IN DIFFERENT TRAITS AND BY THE SIGNIFICANCE OF THE EDUCATIONAL OUTCOMES

Instructor:	*A*	*B*	*C*
Outcome			
(1) Theoretical-mindedness		8,120 10,780	9,380
(2) Interest in teaching	12,740		10,360 11,060
(3) Neurotic tendency	8,900 9,100	7,900	
(4) Prejudices		12,513 9,546	8,514
(5) Knowledge of psychology (paired by initial information)		20,138 19,695	14,210
(6) Knowledge of psychology (paired by intelligence)		19,894	18,879 12,180
Total	30,740	108,582	84,583
Ratio	100	353	275
Total (omitting (6))	30,740	88,692	53,524
Ratio	100	289	174
Total (omitting (5))	30,740	68,753	70,373
Ratio	100	224	229

[1] Perhaps this method of chances may be justified by recalling that "chances of 66 out of 100" means that the event happens 66 times in 100 cases, or is computed to happen 66 times out of 100. One teacher, accordingly, would do better than another in 66 courses out of 100.

A curious feature which deserves some comment is that although B excels his colleagues in effecting changes in information about educational psychology where the classes are paired in that subject by means of a pre-test, C excels in the same way when the contrasted groups are matched for intelligence. The explanation probably lies in the low correlation between the two variables (.33) and in the presumably more uniform accomplishment ratios of C's group which appear in Table III. Of more general significance is the fact that every teacher occupies highest, lowest, and middle ground in some one of the bases of comparison, indicating that teaching skill is highly specialized according to different aspects of pupil development. If one wishes a student to acquire a more stable personality, he would probably move in that direction most readily under A's influence; if interest in teaching needs cultivation, C would appear best for that end; and B would do more to diminish social prejudices than the other two men. In all cases, some great gains in a number of things would be bought at the price of less progress in other respects.

This final table exhausts all the available and defensible objective methods of comparing the three men for variations in instructional skill. If the statement, "By their fruits shall ye know them," be sound philosophy, then the whole doctrine of probability would declare teacher B the best instructor of this course in educational psychology. As an historical and biographical fact, this conclusion appears inevitable. Of course, this fact alone tells us nothing about future efficiency, since the passage of a few years may fully alter the relative ranking. Neither does it suggest anything definite about fitness to teach even a closely related subject, such as industrial psychology or educational measurements. Moreover, the omission of measures for other outcomes of instruction such as skill in applying one's knowledge in "practical" situations is a major defect in the testing program. Performance tests for honesty, the social graces, etc., might have given a clue to some other influential intangibles emanating from the instructor. Unfortunately, in any testing enterprise, one cannot employ all the available measures of conduct and ability without creating a ludicrous and impossible situation; one is always driven to work with a *sampling* of important outcomes. If the sample is unrepresentative of the product of teaching activity, then of course the deduction made above is false. Until that is demonstrated, the decision assigning first rank to B, second to C, and third to A must stand.

CHAPTER V

THE EFFECT OF FORMAL SCHOOL INSTRUCTION UPON PERSONALITY ORGANIZATION

A. *Changes in various attitudes resulting from college courses*

An important by-product of this study is the comforting assurance that attitudes can be altered in approved directions under adequate classroom conditions. It should not be forgotten that all groups, irrespective of the teacher involved, made gains, only some made more than others. Early investigators in this field appear to have been disheartened by the apparent imperviousness of group prejudices to educational influences. Symonds,[1] *e.g.*, tested the liberal-mindedness of students in Hawaiian schools, from the eighth grade through the high school and through the University of Hawaii. He found that the proportion of liberal answers to his questionnaire was remarkably uniform throughout the successive grades of the educational system. Another equally discouraging commentary on college life is the fact that Young (as cited by Allport and Katz[2]) found no significant abatement of prejudice produced by his course on race problems, a course designed largely to accomplish this very purpose. Students memorized intellectually the information presented, but the emotional aspect of their attitudes was largely unaffected. An unpublished master's essay by Walker on the extent of international-mindedness among college students revealed little if any differences between the average freshmen and the average senior at a state university.[3] Summarizing the findings of a test on the opinions of college students, Jones[4] says, "The most significant general educational implication is the slight effect that college training has on the real opinions of students who are now seniors in the Arts and Science college. There is almost as large a proportion in the senior class who are conservative or reactionary in their responses as in the freshman class."

Something must be fundamentally wrong with the teaching organization where such conditions obtain. A happier outcome may

[1] P. M. Symonds, A social attitudes questionnaire, *J. Ed. Psychol.*, 1925, *16*, 316–322.

[2] F. H. Allport and D. Katz, *Students' Attitudes*, Craftsman Press, Syracuse, 1931.

[3] H. C. Walker, Changes in international and interracial attitudes from the freshman to the senior year. The abstract appears in *Studies in Education at the Pennsylvania State College*, Part II, 1932, pages 53–54.

[4] E. S. Jones, Opinions of college students, *J. Applied Psy.*, 1926, 10, 529ff.

be observed in Table VI above, where slight but genuine advances are registered in all concomitant learnings. There are, of course, wide variations in the degree of improvement among the "traits" as the following data show:

TABLE IX

MEAN GAINS IN DIFFERENT TESTS AS A RESULT OF A SEMESTER'S INSTRUCTION IN EDUCATIONAL PSYCHOLOGY (ALL CLASSES COMBINED)

Trait	*Mean Gain*	*S.D.$_{av.}$ of final mean*	*N*	*Chances that no true gain occurred*
Knowledge of educational psychology	26.26	1.95	212	nil
Social prejudices	−11.50	5.58	184	2 in 100
Theoretical-mindedness	1.02	1.15	208	20 in 100
Interest in teaching	22.48	16.60	194	9 in 100
Neurotic tendency	−19.34	9.50	198	2 in 100

This list suggests that some features of personality are more labile and amenable to modification than others, with pure academic information the most plastic and fundamental *Weltanschauung* the most resistant. Presumably, this is because the former is the primary goal of instruction, but the question naturally arises: If so much can be accomplished with the "secondary" objectives when they are pushed into the background, how much more could be attained if they were elevated to the level of consciously and deliberately pursued ends? Under such a scheme of instruction, we should indeed be teaching not just an isolated subject but the pupil in his unique entirety. Some may protest that this is vicious indoctrination and not sound education. But as philosophers and sociologists well know this distinction is a vanishing one. *All* education is indoctrination of a kind, the sole difference being in the nature of the sanctions applied. Knowingly or not, the educational system of America tends to produce good little capitalists just as teaching in the Soviet world is directed toward the formation of good little Communists. Were teachers more generally aware of its inevitability, they would be more deeply concerned with promoting indoctrination of the right type in the light of objectively-determined values.

B. *Changes in the mean valuation types of students in different subjects and under different instructors*

In this section will be found supplementary material showing the shifts in student scores on the Allport-Vernon scale of values blank between the beginning and end of the semester from college

and high-school classes other than those participating in the main body of the experiment. It was thought that the mean valuational profile of each group should be affected in a characteristic manner by the nature of the subject pursued *or* the value organization of the teacher in charge.[5] For instance, one would expect a gain in aesthetic appreciation (as evidenced by a rise in that portion of the scale) after a study of English literature in which canons of taste and standards of beauty in poetry and prose are developed. It is

TABLE X

SEMESTER CHANGES IN SCORES ON DIFFERENT PARTS OF THE ALLPORT-VERNON SCALE OF VALUES BLANK MADE BY STUDENTS IN DIFFERENT COLLEGE CLASSES

	Theo-retical	*Eco-nomic*	*Aes-thetic*	*So-cial*	*Polit-ical*	*Relig-ious*
A. *Literature* class; N = 24						
Instructor's profile (L):	28.5	10.5	41.5	40	19	40.5
Initial Mean	31.33	37.08	23.83	24.79	32.75	30.00
S. D.	4.65	6.36	5.66	6.03	6.37	8.67
Final Mean	31.75	36.79	23.25	25.54	31.96	30.33
S. D.	6.34	5.07	4.52	5.64	6.00	6.07
Differences	.42	−.29	−.58	.75	−.79	.33
B. *Sociology* class; N = 24						
Instructor's profile (W):	39	33.5	25.5	32.5	34.5	15
Initial Mean	27.92	28.33	32.42	32.46	29.08	33.29
S. D.	6.98	7.31	10.82	7.84	8.73	10.77
Final Mean	28.00	29.71	31.21	32.08	28.83	31.88
S. D.	6.27	7.57	9.70	6.80	8.54	9.67
Differences	.08	1.38	−1.21	−.38	−.25	−1.41
C. *Political Science* class; N = 29						
Instructor's profile (A):	36	17	45	26	26	30
Initial Mean	32.52	31.28	26.31	28.31	32.93	28.21
S. D.	6.55	5.84	8.24	5.36	7.10	7.89
Final Mean	31.93	31.07	25.90	29.90	35.24	25.66
S. D.	7.87	4.70	7.71	7.15	5.67	8.31
Differences	−.59	−.21	−.41	1.59	2.31	−2.55
D. *Business Psychology* class; N = 36						
Instructor's profile (H):	40	31	23	32	21	25
Initial Mean	32.94	33.61	22.39	27.64	32.78	31.63
S. D.	7.03	5.89	6.99	5.58	6.40	7.43
Final Mean	33.76	31.56	25.39	27.14	32.36	29.89
S. D.	6.14	6.62	7.44	5.47	7.14	7.33
Differences	.82	−2.05	3.00	−.50	−.42	−1.74

[5] The only remotely similar endeavor in this direction is a quasi-experimental monograph by W. O. Döring (*Untersuchungen zur Psychologie des Lehrers*, Leipzig, Quelle und Meyer, 1925). Applying Spranger's criteria, he identifies Pestalozzi as the social type of teacher, Humboldt as the aesthetic, Socrates as the theoretical, Basedow as the economic, Francke as the religious, and the representative Jesuit pedagogues as the political.

also conceivable that an instructor with a strong "social" trend will subtly increase that dispositional bias in his pupils through the daily expression of opinions, gesture, and other mild hints by which personal viewpoint is revealed.

The data obtained were disappointing, since no consistent trend appeared. Table X shows the results in four college sophomore classes in literature, sociology, political science, and business psychology. The most casual glance indicates the absence of any sort of uniformity. What shall one make out of a situation in which a study of literature makes one less aesthetically-minded, a course in sociology less "social," and a semester of business psychology less inclined to prize economic ends? Certainly the largest displacement—a gain of 3.00 in aesthetic attitude after work in business psychology—is utterly baffling.

The absence of any association between the character of the course pursued and the type of attitudinal change resulting makes one suspect that the instructor's own mental constitution may be a more potent influence than the material studied. An inspection of Table X lends no support to this suspicion. Three Junior classes

TABLE XI

CHANGES IN THE VARIOUS DIVISIONS OF THE ALLPORT-VERNON TEST IN THREE HIGH-SCHOOL CLASSES IN AMERICAN HISTORY TAUGHT BY DIFFERENT TEACHERS

	Theo-retical	*Eco-nomic*	*Aes-thetic*	*So-cial*	*Polit-ical*	*Relig-ious*
Section 1: N = 24						
Instructor's profile (C):	34	36	23	26	30	31
Initial Mean	29.63	28.0	23.63	28.92	31.25	37.63
S. D.	5.38	4.70	6.74	6.07	4.88	7.13
Final Mean	30.50	29.00	23.38	27.70	31.33	37.54
S. D.	5.01	4.70	5.93	5.07	4.04	6.47
Differences	.87	1.00	−.25	−1.22	.08	−.09
Section 2: N = 30						
Instructor's profile (H):	27	19	24	33	40	37
Initial Mean	29.00	28.33	25.80	29.57	31.03	37.47
S. D.	5.35	5.82	6.40	5.22	5.98	5.97
Final Mean	29.10	28.10	27.13	28.30	30.43	36.67
S. D.	6.56	6.90	6.41	6.25	5.94	7.16
Differences	.10	−.23	1.33	−1.27	−.60	−.80
Section 3: N = 23						
Instructor's profile (S):	33	34	16	36	26	35
Initial Mean	25.26	28.70	29.70	28.83	32.30	34.83
S. D.	6.50	6.69	6.08	6.00	4.26	7.77
Final Mean	26.43	28.30	28.60	29.42	31.65	35.74
S. D.	7.82	4.95	7.38	6.21	9.25	7.64
Differences	1.17	−.40	−.90	.59	−.65	.91

in American history in a local high school, taught by three different men, were examined for further evidence on this point. The data are presented in Table XI.

C's dominant value lies in the *economic* field and the maximum mean gains among his pupils occur in his two highest values and the major losses in his two lowest. But H, an outspoken *Politiker,* diminishes rather than adds to his students' esteem of power, while in his next weakest trait—the aesthetic—they make the greatest increase! S's group generally moves toward his dominant values and away from his lowest; the greatest irregularity occurs in the *economic* column.

Too few instructors are involved to reach any conclusion and the data reveal a highly uneven picture. No relations stand out clearly. It seems that students are repelled as often as they are attracted by the axiological organization of their teachers' personalities; or perhaps they appreciate different attitudes in a detached intellectual way but remain emotionally unaffected thereby. A further explanation may be that the Allport test definitely exposes fundamental and enduring attitudes of the sort represented by Shand's sentiments, which are established so firmly at an early age that modification is difficult or impossible. At any rate, the materials of Tables X and XI help one understand why differences in "theoretical-mindedness" as measured by this scale were the smallest and least certain of the changes registered by the instructors analyzed in this report.

CHAPTER VI

CONCLUSIONS

Can teaching efficiency be measured? If measurement implies a maximum amount of objectivity, the answer of this investigation is an emphatic "Yes." How accurately can teaching efficiency be measured? One must pause before giving a reply. A reasonable claim would be that a small number of teachers of the same subject can be fairly precisely arranged in a serial order of merit, but that the exact magnitude of the gap between any two individual positions can be only roughly approximated. Nevertheless, to be able to rank teachers on the basis of objectively ascertainable changes in their pupils by procedures which take account of *all* educational objectives is eminently worth while. The prediction is not unwarranted that the most promising future methods of investigating instructional skill will attempt to determine the amount and kind of personality modifications transpiring between initial and final determinations of status when subject to a given variety of human environment. The testing program should undoubtedly be expanded just as surely as the mathematical technique ought to be refined, but the broad outlines of this type of experimental approach ought to be preserved until supplanted by a more direct measuring instrument.

The shortcomings of this procedure are unfortunately serious. Its chief limitation is the impracticability of employing it on any large scale. The demands upon the time, energy, and skill of a well-trained educational expert are probably more than the ordinary community school system could afford. In addition, the method as illustrated in this account is applicable only to those circumstances where classes consisting of similar pupils are pursuing the same subjects. Where these conditions are met, however, the scheme seems worthy of further trial.

CHAPTER VII

SUMMARY

Three instructors of educational psychology, A, B and C, who taught classes comprising 55, 59, and 105 students, respectively, cooperated in an investigation of their relative teaching efficiency by administering tests of achievement, attitudes, and personality traits at both the beginning and end of a semester. Individuals in all three groups were then paired according to pre-test scores in each measure and the differences in average gain under various teachers determined. The basic assumption behind the statistical comparison was that the most effective instructor was the one who was responsible for the greatest number of desirable changes in the pupils. With this as a standard and after application of a system of weighting to the objective results according to the importance of the outcomes, an order of efficiency was finally determined. None of the differences between the groups met the conventional requirements of absolute reliability, although the probabilities were heavily in favor of the order found. Each teacher considered showed some superiority where his colleagues exhibited relative weaknesses, implying a high degree of specificity in pedagogical talent. Teaching ability apparently can be measured if one accepts the propriety of summating separate excellences with the result that a mean plane of instructional skill emerges. The best teacher will then be the one with the highest average level of success in producing an array of valuable personality modifications in the human beings under his influence. He can be identified with reasonable accuracy by means of an extensive test program wherever he is teaching a course or subject in which other instructors are engaged.

R.W.B. JACKSON LIBRARY
3 0005 03037621 7

150.8
A673
no. 154
1933
Archives of psychology

150.8
A673
no. 154
1933
Archives of psychology

79. Memory Value of Advertisements: E. R. Brandt. $1.25.
80. Critical Examination of Test-Scoring Methods: R. G. Anderson. $1.00.
81. Thermal Discrimination and Weber's Law: E. A. K. Culler. $1.75.
82. Correlational Analysis of Typing Proficiency: L. Ackerson. $1.50.
83. Recall as a Function of Perceived Relations: C. B. Key. $1.25.
84. Consistency of Rate of Work: C. E. Dowd. $1.00.
85. Experimental Investigation of Recovery from Work: S. L. Crawley. $1.25.
86. Facilitation and Inhibition: T. N. Jenkins. $1.00.
87. Variability of Performance in the Curve of Work: J. D. Weinland. $1.00.
88. Mental Hygiene Inventory: S. D. House. $1.50.
89. Mental Set and Shift: A. T. Jersild. $1.25.
90. Experimental Investigation of Rest Pauses: C. W. Manzer. $1.25.
91. Routine and Varying Practice as Preparation for Adjustment to a New Situation: L. W. Crafts. $1.00.
93. Speed and Other Factors in "Racial" Differences: O. Klineberg. $1.50.
94. Relation of Reaction Time to Intelligence, Memory, and Learning: V. W. Lemmon. 80c.
95. Is the Latent Time in the Achilles Tendon Reflex a Criterion of Speed in Mental Reactions? G. H. Rounds. $1.25.
96. Predictive Value of Tests of Emotional Stability Applied to College Freshmen: E. G. Flemming. $1.00.
97. Vocabulary Information Test: A. L. Weeks. $1.00.
98. Effect of Temporal Arrangements of Practice on the Mastery of an Animal Maze: S. A. Cook. 80c.
99. Recognition Time as a Measure of Confidence: G. H. Seward. $1.00.
100. Precision and Accuracy: G. W. Hartman. 80c.
101. Group Test of Home Environment: E. M. Burdick. $1.50.
102. Effect of Material on Formal Syllogistic Reasoning: M. C. Wilkins. $1.25.
103. Effect of Incentives on Accuracy of Discrimination: H. C. Hamilton. $1.25.
104. Validity of Norms with Special Reference to Urban and Rural Groups: M. E. Shimberg. $1.25.
105. Blood Pressure Changes in Deception: M. N. Chappell. 80c.
106. Experimental Comparison of Psychophysical Methods: W. N. Kellogg. $1.25.
107. Measurement of Verbal and Numerical Abilities: M. M. R. Schneck. $1.00.
108. Perseverative Tendency in Pre-School Children. A Study in Personality: H. M. Cushing. $1.00.
109. Preliminary Study of the Effect of Training in Junior High School Shop Courses: L. D. Anderson. 80c.
110. Music Appreciation: M. J. Adler. $1.50.
111. Motivation in Fashion: E. B. Hurlock. $1.00.
112. Equality Judgments in Psychophysics: W. N. Kellogg. $1.00.
113. Illusions in the Perception of Short Time Intervals: N. Israeli. 80c.
114. Further Studies of the Reading-Recitation Process in Learning: Skaggs, Grossman, Krueger & Krueger. 80c.
115. Factors Affecting the Galvanic Reflex: R. C. Davis. $1.00.
116. Infant's Feeding Reactions During the First Six Months: R. Ripin. 80c.
117. Measurement of Mental Deterioration: H. Babcock. $1.25.
118. Phenomenon of Postural Persistence: L. S. Selling. $1.00.
119. American Council on Education Rating Scale: F. F. Bradshaw. $1.00.
120. Group Factor in Immediate Memory: A. Anastasi. $1.00.
121. Individual Differences in the Sense of Humor and Temperamental Differences: P. Kambouropoulou. $1.00.
122. Suggestibility in Normal and Hypnotic States: G. W. Williams. $1.00.
123. Analytical Study of the Conditioned Knee-Jerk: G. R. Wendt. $1.25.
124. Race Differences in the Organization of Numerical and Verbal Abilities: J. W. Dunlap. $1.25.
125. Errors of Measurement and Correlation: E. E. Cureton. $1.25.
126. Experience Factors, Test Scores and Efficiency of Women Office Workers: N. Bird. $1.00.
127. Delayed Reactions of Infants: C. N. Allen. 80c.
128. Factors Measured by the Thorndike Intelligence Examination for H. S.; J. G. Peatman. $1.00.
129. Educational Success and Failure in Supernormal Children: J. Regensburg. $1.75.
130. Effect of Practice on Visual Perception of Form: J. P. Seward. $1.00.
131. Relation to College Grades of Some Factors other than Intelligence: D. Harris. 80c.
132. Study of Psychological Differences Between "Racial" and National Groups in Europe: O. Klineberg. $1.00.
133. Emotional Differences of Delinquent and Non-Delinquent Girls of Normal Intelligence: A. Courthial. $1.25.
134. Learning and Retention of Pleasant and Unpleasant Activities: H. Cason. $1.25.
135. Investigation of Brightness Constancy: R. B. MacLeod. $1.25.
136. The Rorschach Test Applied to Feeble-Minded Group: S. J. Beck. $1.00.
137. Retention after intervals of Sleep and of Waking: E. B. van Ormer. $1.00.
138. Stimulus Temperature and Thermal Sensation: F. Heiser. $1.00.
139. Energy Cost Measurements on Curve of Work: H. J. Schubert. $1.00.
140. Technique for the Measurements of Attitudes: R. Likert. 80c.
141. Speed Factor in Mental Tests: P. H. DuBois. 80c.
142. Further Studies on the Memory Factor: A. Anastasi. $1.00.
143. An Experimental Study on Variability of Learning: S. E. Asch. $1.00.
144. Development of Inventory for Measurement of Inferiority Feelings at H. S. Level: R. B. Smith. $1.50.
145. The Psychological Effects of Oxygen Deprivation: R. A. McFarland. $1.50.
146. Relation of Subliminal to Supraliminal Learning: O. A. Simley. 80c.
147. Effects of Noise upon Certain Psychological and Physiological Processes: F. L. Harmon. $1.25.
148. Conditioned Responses in Children: G. S. Razran. $1.50.
149. Resemblance of Parents and Children in General Intelligence: M. C. Outhit. $1.00.
150. Influence of Oral Propaganda Material . . . : K-C. Chen. 80c.
151. Negative or Withdrawal Attitude: H. Pallister. 80c.
152. Judgment in Absolute Units as a Psychophysical Method: J. Bressler. $1.00.
153. Visual Illusions in the Chick: C. N. Winslow. $1.25.
154. Measuring Teaching Efficiency Among College Instructors: G. W. Hartmann. 80c.

DIRECTORY OF
AMERICAN PSYCHOLOGICAL PERIODICALS

AMERICAN JOURNAL OF PSYCHOLOGY—Ithaca, N. Y.; Cornell University.
Subscription $6.50. 624 pages annually. Edited by M. F. Washburn, K. M. Dallenbach, Madison Bentley, and E. G. Boring. Quarterly. General and experimental psychology. Founded 1887.

JOURNAL OF GENETIC PSYCHOLOGY—Worcester, Mass.; Clark University Press.
Subscription $14.00 per year; $7.00 per volume. 1000 pages annually. (2 volumes.) Edited by Carl Murchison. Quarterly. Child behavior, animal behavior, and comparative psychology. Founded 1891.

PSYCHOLOGICAL REVIEW—Princeton, N. J.; Psychological Review Company.
Subscription $5.50. 540 pages annually. Edited by Howard C. Warren. Bimonthly. General psychology. Founded 1894.

PSYCHOLOGICAL MONOGRAPHS—Princeton, N. J.; Psychological Review Company.
Subscription $6.00 per volume. 500 pages. Edited by Herbert S. Langfeld. Published without fixed dates, each issue one or more researches. Founded 1895.

PSYCHOLOGICAL INDEX—Princeton, N. J.; Psychological Review Company.
Subscription $4.00. 300–400 pages. Edited by Walter S. Hunter and R. R. Willoughby. An annual bibliography of psychological literature. Founded 1895.

PSYCHOLOGICAL BULLETIN—Princeton, N. J.; Psychological Review Company.
Subscription $6.00. 720 pages annually. Edited by Edward S. Robinson. Monthly (10 numbers). Psychological literature. Founded 1904.

ARCHIVES OF PSYCHOLOGY—New York, N. Y.; Columbia University.
Subscription $6.00. 500 pages per volume. Edited by R. S. Woodworth. Without fixed dates, each number a single experimental study. Founded 1906.

JOURNAL OF ABNORMAL AND SOCIAL PSYCHOLOGY—Eno Hall, Princeton, N. J.; American Psychological Association.
Subscription $5.00. 448 pages annually. Edited by Henry T. Moore. Quarterly. Abnormal and social. Founded 1906.

PSYCHOLOGICAL CLINIC—Philadelphia, Pa.; Psychological Clinic Press.
Subscription $3.00. 288 pages. Edited by Lightner Witmer. Without fixed dates (9 numbers). Orthogenics, psychology, hygiene. Founded 1907.

PSYCHOANALYTIC REVIEW—Washington, D. C.; 3617 10th St., N. W.
Subscription $6.00. 500 pages annually. Edited by W. A. White and S. E. Jelliffe. Quarterly. Psychoanalysis. Founded 1913.

JOURNAL OF EXPERIMENTAL PSYCHOLOGY—Princeton, N. J.
Psychological Review Company, 700 pages annually. Experimental. Subscription $7.00. Bi-monthly. Edited by S. W. Fernberger. Founded 1916.

JOURNAL OF APPLIED PSYCHOLOGY—Athens, Ohio.
Subscription $5.50. 400 pages annually. Edited by James P. Porter. Bi-monthly. Founded 1917.

JOURNAL OF COMPARATIVE PSYCHOLOGY—Baltimore, Md.; Williams & Wilkins Company.
Subscription $5.00 per volume of 450 pages. Three volumes every two years. Edited by Knight Dunlap and Robert M. Yerkes. Founded 1921.

COMPARATIVE PSYCHOLOGY MONOGRAPHS—Baltimore, Md.; The Johns Hopkins Press.
Subscription $5.00. 400 pages per volume. Knight Dunlap, Managing Editor. Published without fixed dates, each number a single research. Founded 1922.

GENETIC PSYCHOLOGY MONOGRAPHS—Worcester, Mass.; Clark University Press.
Subscription $14.00 per year; $7.00 per volume. 1000 pages annually. (2 volumes.) Edited by Carl Murchison. Monthly. Each number one complete research. Child behavior, animal behavior, and comparative psychology. Founded 1925.

PSYCHOLOGICAL ABSTRACTS—Eno Hall, Princeton, N. J.; American Psychological Association.
Subscription $6.00. 700 pages annually. Edited by Walter S. Hunter and R. R. Willoughby. Monthly. Abstracts of psychological literature. Founded 1927.

JOURNAL OF GENERAL PSYCHOLOGY—Worcester, Mass.; Clark University Press.
Subscription $14.00 per year; $7.00 per volume. 1000 pages annually. (2 volumes.) Edited by Carl Murchison. Quarterly. Experimental, theoretical, clinical, and historical psychology. Founded 1927.

JOURNAL OF SOCIAL PSYCHOLOGY—Worcester, Mass.; Clark University Press.
Subscription $7.00. 500 pages annually. Edited by John Dewey and Carl Murchison. Quarterly. Political, racial, and differential psychology. Founded 1929.